Online Appendix to "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice"

Toru Kitagawa* and Aleksey Tetenov†

November 27, 2017

D Extensions

D.1 Empirical Welfare Maximization with a Capacity Constraint

This section shows a proof of the claim given in Remark 2.1 of the main text that says the expected welfare of $\hat{G}^K$ converges to the maximum at least at $n^{-1/2}$ rate. The result is analogous to Theorem 2.1, with the additional term corresponding to potential welfare losses due to estimation errors of $P_X(G)$.

Theorem D.1. *Under Assumption 2.1,*

$$\sup_{P\in\mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G\in\mathcal{G}} W^K(G) - W^K(\hat{G}^K)\right] \leq C_1\frac{M}{\kappa}\sqrt{\frac{v}{n}} + C_1\frac{M}{K}\sqrt{\frac{v}{n}},$$

where C_1 is the universal constant in Lemma A.4.

Proof. Since $W^K(G) - W^K(G') = V^K(G) - V^K(G')$ for all G, G',

$$\sup_{P\in\mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G\in\mathcal{G}} W^K(G) - W^K(\hat{G}^K)\right] = \sup_{P\in\mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G\in\mathcal{G}} V^K(G) - V^K(\hat{G}^K)\right], \tag{D.1}$$

and we focus on bounding the latter expression.

Since $\hat{G}^K$ maximizes $V_n^K(G)$, $V_n^K(\tilde{G}) \leq V_n^K(\hat{G}^K)$ for any $\tilde{G}\in\mathcal{G}$ and

$$\begin{aligned} V^K(\tilde{G}) &\leq V_n^K(\tilde{G}) + \sup_{G\in\mathcal{G}}\left|V_n^K(G) - V^K(G)\right| \\ &\leq V_n^K(\hat{G}^K) + \sup_{G\in\mathcal{G}}\left|V_n^K(G) - V^K(G)\right| \\ &\leq V^K(\hat{G}^K) + 2\sup_{G\in\mathcal{G}}\left|V_n^K(G) - V^K(G)\right|. \end{aligned}$$

Applying the inequality for all $\tilde{G}\in\mathcal{G}$, we obtain

$$\sup_{G\in\mathcal{G}} V^K(G) - V^K(\hat{G}^K) \leq 2\sup_{G\in\mathcal{G}}\left|V_n^K(G) - V^K(G)\right|,$$

*Cemmap/University College London, Department of Economics. Email: t.kitagawa@ucl.ac.uk
†University of Bristol, Email: a.tetenov@bristol.ac.uk

which is also true in expectation over P^n.

The welfare gain estimation error for any treatment rule G could be bounded from above by:

$$\begin{aligned}
&\left|V_n^K(G) - V^K(G)\right| = \left|\frac{K}{\max\{K, P_{X,n}(G)\}} \cdot V_n(G) - \frac{K}{\max\{K, P_X(G)\}} \cdot V(G)\right| \\
&\leq \frac{K}{\max\{K, P_{X,n}(G)\}} \cdot |V_n(G) - V(G)| + V(G) \cdot \left|\frac{K}{\max\{K, P_{X,n}(G)\}} - \frac{K}{\max\{K, P_X(G)\}}\right| \\
&\leq |V_n(G) - V(G)| + \frac{M}{K} \cdot |P_{X,n}(G) - P_X(G)|.
\end{aligned}$$

The second line comes from subtracting and adding $\frac{K}{\max\{K,P_{X,n}(G)\}}V(G)$ and then applying the triangle inequality. The third line uses inequalities $\frac{K}{\max\{K,P_{X,n}(G)\}} \leq 1$ and $V(G) \leq M$ (from Assumption 2.1 (BO)), and the observation that for any $a, b \in \mathbb{R}$ and $c > 0$,

$$\left|\frac{c}{\max\{c,a\}} - \frac{c}{\max\{c,b\}}\right| = \left|\frac{c(\max\{c,b\} - \max\{c,a\})}{\max\{c,a\} \cdot \max\{c,b\}}\right| \leq \frac{|\max\{c,b\} - \max\{c,a\}|}{c} \leq \frac{|b-a|}{c}.$$

Then

$$\begin{aligned}
&\sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} V^K(G) - V^K(\hat{G}^K)\right] \leq 2 \sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} \left|V_n^K(G) - V^K(G)\right|\right] \\
&\leq 2 \sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} |V_n(G) - V(G)|\right] + 2\frac{M}{K} \sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} |P_{X,n}(G) - P_X(G)|\right]
\end{aligned}$$

Note that since the class $\mathcal{G}$ has VC-dimension $v < \infty$, the classes of functions

$$\begin{aligned}
f_G(Y, D, X) &\equiv \left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right) \cdot 1\{X \in G\}, \\
h_G(Y, D, X) &\equiv 1\{X \in G\} - 1/2,
\end{aligned}$$

are VC-subgraph classes with VC-dimension no greater than v by Lemma A.1. These classes of functions are uniformly bounded by $M/(2\kappa)$ and $1/2$. Since $V_n(G) = E_n(f_G)$, $V(G) = E_P(f_G)$, $P_{X,n}(G) = E_n(h_G) + 1/2$ and $P_X(G) = E_P(h_G) + 1/2$, we could apply Lemma A.4 and obtain

$$\sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} V^K(G) - V^K(\hat{G}^K)\right] \leq C_1 \frac{M}{\kappa}\sqrt{\frac{v}{n}} + C_1 \frac{M}{K}\sqrt{\frac{v}{n}}.$$

The theorem's result follows from (D.1). □

D.2 Demeaned EWM

Define the demeaned population welfare as

$$W^{dm}(G) \equiv W(G) - E_P[Y],$$

then $\sup_{G\in\mathcal{G}} W^{dm}(G) = \sup_{G\in\mathcal{G}} W(G) - E_P[Y] = W^*_{\mathcal{G}} - E_P[Y]$. Analogously to (2.2), for any $\tilde{G}\in\mathcal{G}$,

$$W^{dm}(\tilde{G}) - W^{dm}(\hat{G}^{dm}_{EWM}) \leq 2\sup_{G\in\mathcal{G}}\left|W^{dm}_n(G) - W^{dm}(G)\right|,$$

therefore

$$W^*_{\mathcal{G}} - W(\hat{G}^{dm}_{EWM}) \leq 2\sup_{G\in\mathcal{G}}\left|W^{dm}_n(G) - W^{dm}(G)\right|.$$

Note that since $Y^{dm}_i = Y_i - E_n[Y_i]$,

$$\begin{aligned} W^{dm}_n(G) &= E_n\left[\frac{Y^{dm}_i D_i}{e(X_i)}\cdot 1\{X_i\in G\} + \frac{Y^{dm}_i(1-D_i)}{1-e(X_i)}\cdot 1\{X_i\notin G\}\right] \\ &= W_n(G) - E_n[Y_i]\cdot E_n\left[\frac{D_i}{e(X_i)}\cdot 1\{X_i\in G\} + \frac{1-D_i}{1-e(X_i)}\cdot 1\{X_i\notin G\}\right], \end{aligned}$$

and since $|E_n(Y_i)|\leq M/2$,

$$\begin{aligned} \left|W^{dm}_n(G) - W^{dm}(G)\right| &\leq |W_n(G) - W(G)| \\ &+ \left|E_n[Y_i]\cdot E_n\left[\frac{D_i}{e(X_i)}\cdot 1\{X_i\in G\} + \frac{1-D_i}{1-e(X_i)}\cdot 1\{X_i\notin G\}\right] - E_P[Y]\right| \\ &\leq |W_n(G) - W(G)| \\ &+ |E_n(Y_i) - E_P[Y]| \\ &+ \frac{M}{2}\cdot\left|E_n\left[\frac{D_i}{e(X_i)}\cdot 1\{X_i\in G\} + \frac{1-D_i}{1-e(X_i)}\cdot 1\{X_i\notin G\}\right] - 1\right|. \end{aligned}$$

Similarly to the proof of Theorem 2.1, Lemma A.4 applies to all three terms with envelopes $M/(2\kappa)$, $M/2$, and $M/(2\kappa)$, thus

$$E_{P^n}\left[W^*_{\mathcal{G}} - W(\hat{G}^{dm}_{EWM})\right] \leq 2E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|W^{dm}_n(G) - W^{dm}(G)\right|\right] \leq C_1 M\left(\frac{2}{\kappa}+1\right)\sqrt{\frac{v}{n}}.$$

D.3 Multiple Treatments

It is feasible to extend the current approach to situations with multiple treatments. Suppose there are K treatments denoted by $D\in\{1,\ldots,K\}$. Let $e_k(x) = P(D=k|X=x)$, $k=1,\ldots,K$, be the propensity scores in the experimental data, and $\{Y_k : k=1,\ldots,K\}$ be the potential outcomes for each treatment. Define a treatment assignment policy by a K-partition of the covariate space $\mathcal{X}$, $\mathbf{G} = (G_1,\ldots,G_K)$, where $G_1,\cdots,G_K\subset\mathcal{X}$ are non-intersecting subsets that partition $\mathcal{X}$ into K regions. For each $k=1,\ldots,K$, G_k specifies a subpopulation to which treatment $D=k$ is assigned.

Under unconfoundedness, $(Y_1, \ldots, Y_K) \perp D|X$, consider the following empirical welfare criterion;

$$W_n(\mathbf{G}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{Y_i \cdot 1\{D_i = k\}}{e_k(X_i)} \cdot 1\{X_i \in G_k\},$$

which unbiasedly estimates the population welfare attained by policy $\mathbf{G}$,

$$W(\mathbf{G}) = \sum_{k=1}^{K} E[Y_k \cdot 1\{X \in G_k\}].$$

Consider setting the space of policies to $\mathbb{G} = \{\mathbf{G} : G_1 \in \mathcal{G}, \ldots, G_K \in \mathcal{G}, \mathbf{G} \text{ partitions } \mathcal{X}\}$, where $\mathcal{G}$ is a VC-class of subsets in $\mathcal{X}$ including $\emptyset$ such that K distinct subsets in $\mathcal{G}$ can form a partition of $\mathcal{X}$. For instance, when $\mathcal{X} = \mathbb{R}$, a class of connected intervals of the form $\mathcal{G} = \{(x, x'] : -\infty \leq x \leq x' \leq \infty\} \cup \emptyset$ is a VC-class that allows us to pick K-distinct subsets partitioning $\mathbb{R}$. The EWM rule can be then obtained as $\hat{\mathbf{G}}_{EWM} \in \arg\max_{\mathbf{G}\in\mathbb{G}} W(\mathbf{G})$.

Analogous to derivation of inequality (2.3) in the paper, we can bound the welfare loss of the EWM rule as

$$\sup_{\mathbf{G}\in\mathbb{G}} W(\mathbf{G}) - W(\hat{\mathbf{G}}_{EWM}) \leq \sum_{k=1}^{K} 2 \sup_{G_k \in \mathcal{G}} \left| W_n^k(G_k) - W^k(G_k) \right|,$$

where $W_n^k(G_k) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \cdot 1\{D_i=k\}}{e_k(X_i)} \cdot 1\{X_i \in G_k\}$ and $W^k(G_k) \equiv E[Y_k \cdot 1\{X \in G_k\}]$. Assuming bounded outcomes $Y \in [-M/2, M/2]$ and strict overlap, in the sense that $e_k(x) \in [\kappa, 1-\kappa]$ for all x and $k = 1, \ldots, K$ for some $\kappa > 0$, we apply Lemmas A.1 and A.4 to obtain the mean of $\sup_{G_k\in\mathcal{G}} \left| W_n^k(G_k) - W^k(G_k) \right|$ bounded from above by $C_1 M\sqrt{v/n}/\kappa$. Hence, the whole welfare loss can be bounded from above by that of Theorem 2.1 multiplied by the number of treatments K.

Computing $\hat{\mathbf{G}}_{EWM}$ presents additional challenges when the EWM framework is extended from binary to multiple treatment case. We leave an investigation of computational procedures in this setting for future research.

D.4 Comparison with the Nonparametric Plug-in Rule

The plug-in treatment choice rule (1.13) with parametrically or nonparametrically estimated $m_1(x)$ and $m_0(x)$ is intuitive and simple to implement. In situations where flexible treatment assignment rules are allowed and the dimension of conditioning covariates is small, the nonparametric plug-in rule would be a competing alternative to the EWM approach. In this section, we review the welfare loss convergence rate results of the nonparametric plug-in rule and discuss potential advantages and disadvantages of these two approaches.

We denote the class of data generating processes that satisfy Assumptions 2.1 (UCF), (BO), (SO), Assumption 2.2 (MA), and Assumption E.1 by $\mathcal{P}_{smooth}(M, \kappa, \alpha, \eta, \beta_m)$. Given the smoothness assumption of the regression equations, we consider estimating m_1 and m_0 by local polynomial estimators of degree $(\beta_m - 1)$. The convergence rate results of the nonparametric plug-in classifiers shown in Theorem 3.3 of Audibert and Tsybakov (2007) can be straightforwardly extended to the treatment choice context, resulting in

$$\sup_{P \in \mathcal{P}_{smooth}(M,\kappa,\alpha,\eta,\beta_m)} E_{P^n}\left[W(G^*_{FB}) - W(\hat{G}_{plug-in})\right] \leq O\left(n^{-\frac{1+\alpha}{2+d_x/\beta_m}}\right). \tag{D.2}$$

Furthermore, if $\alpha\beta_m \leq d_x$, Theorem 3.5 of Audibert and Tsybakov (2007) applied to the current treatment choice setup shows that the nonparametric plug-in rule attains the rate lower bound i.e., for any treatment rule $\hat{G}$,

$$\sup_{P \in \mathcal{P}_{smooth}(M,\kappa,\alpha,\eta,\beta_m)} E_{P^n}\left[W(G^*_{FB}) - W(\hat{G})\right] \geq O\left(n^{-\frac{1+\alpha}{2+d_x/\beta_m}}\right)$$

holds.

In practically relevant situations where $\alpha\beta_m \leq d_x$,[1] a naive comparison of the welfare loss convergence rate of the plug-in rule presented here with that of EWM (Theorems 2.3 and 2.4) would suggest that in terms of the welfare loss converge rate, the EWM rule would outperform the nonparametric plug-in rule. It is, however, important to notice that the classes of data generating processes over which the uniform rates are ensured differ between the two cases. $\mathcal{P}_{smooth}(M, \kappa, \alpha, \eta, \beta_m)$ is constrained by smooth regression equations and continuously distributed X, whereas $\mathcal{P}_{FB}(M, \kappa, \alpha, \eta)$ considered in Theorems 2.3 and 2.4 allows for discontinuous regression equations and no restriction on the marginal distribution of X's. Assumption 2.2 (FB) on $\mathcal{P}_{FB}(M, \kappa, \alpha, \eta)$ requires that $\{x : \tau(x) \geq 0\}$ belongs to the pre-specified VC-class $\mathcal{G}$, whereas $\mathcal{P}_{smooth}(M, \kappa, \alpha, \eta, \beta_m)$ is free from such assumption. This non-nested relationship between $\mathcal{P}_{FB}(M, \kappa, \alpha, \eta)$ and $\mathcal{P}_{smooth}(M, \kappa, \alpha, \eta, \beta_m)$ makes the naive rate comparison between (D.2) and Theorem 2.3 less meaningful because a data generating process in $\mathcal{P}_{smooth}(M, \kappa, \alpha, \eta, \beta_m)$ that yields the slowest convergence rate for the nonparametric plug-in rule is in fact excluded from $\mathcal{P}_{FB}(M, \kappa, \alpha, \eta)$. Accordingly, unless we can assess which one of $\mathcal{P}_{smooth}(M, \kappa, \alpha, \eta, \beta_m)$ and $\mathcal{P}_{FB}(M, \kappa, \alpha, \eta)$ is more

[1] In an analogy to the Proposition 3.4 of Audibert and Tsybakov (2007), when the class of data generating processes is assumed to have $\alpha\beta_m > d_x$, no data generating process in this class can have the conditional treatment effect $\tau(x) = 0$ in an interior of the support of P_X. In the practice of causal inference, we a priori would not restrict the plausible data generating processes only to these extreme cases; therefore, the class of data generating processes with $\alpha\beta > d_x$ would be less relevant in practice.

likely to contain the true data generating process, these rate results offer us limited guidance on the procedure that should be used in a given application.

In practical terms, we consider these two distinct approaches as complementary, and our choice between them should be based on available assumptions and the dimension of covariates in a given application. With knowledge of the propensity score, a practical advantage of the EWM rule is that the welfare loss convergence rate does not directly depend on the dimension of X, so when an available credible assumption on the level set $\{x : \tau(x) \geq 0\}$ implies a certain class of decision sets with a finite VC-dimension, the EWM approach offers a practical solution to get around the curse of dimensionality of X. A potential drawback of using the EWM rule is the risk of misspecification of $\mathcal{G}$, i.e., if Assumption 2.2 (FB) is not valid, the EWM rule only attains the second-best welfare, whereas the nonparametric plug-in rule is guaranteed to yield the first-best welfare in the limit. Another aspect of comparison is that the performance of the EWM rule is stable regardless of whether the underlying data generating processes, including the marginal distribution of X and the regression equations $m_1(X)$ and $m_0(X)$, are smooth or not. In terms of implementation, the EWM approach becomes particularly attractive when the class of candidate decision sets $\mathcal{G}$ is given exogenously, since the user does not have to specify any smoothing parameter in this case. In contrast, when the user can freely choose $\mathcal{G}$, the welfare performance of the EWM rule can be sensitive to how to choose $\mathcal{G}$, similarly to that the performance of nonparametric plug-in rule can be sensitive to the choice of the smoothing parameter.

E Hybrid EWM with Local Polynomial Estimators

This section focuses on the hybrid EWM approaches with local polynomial estimators for $\tau(x)$ and $e(x)$,. We spell out classes of data generating processes $\mathcal{P}_m$ and $\mathcal{P}_e$ as well as ψ_n, $\tilde{\psi}_n$, ϕ_n, and $\tilde{\phi}_n$ that satisfy Condition 2.1 and the assumption of Theorem 2.6.

E.1 Assumptions, Estimators, and Welfare Convergence Rates

Consider the m-hybrid approach in which the leave-one-out local polynomial estimators are used to estimate $m_1(X_i)$ and $m_0(X_i)$, i.e., $\hat{m}_1(X_i)$ and $\hat{m}_0(X_i)$ are constructed by fitting the local polynomials excluding the i-th observation.[2] For any multi-index $s = (s_1, \ldots, s_{d_x}) \in \mathbb{N}^{d_x}$ and any $(x_1, \ldots, x_{d_x}) \in \mathbb{R}^{d_x}$, we define $|s| \equiv \sum_{i=1}^{d_x} s_i$, $s! \equiv s_1! \cdots s_{d_x}!$, $x^s \equiv x_1^{s_1} \cdots x_{d_x}^{s_{d_x}}$, and $\|x\| \equiv$

[2]The reason to consider the leave-one-out fitted values is to simplify analytical verification of Condition 2.1. We believe that the welfare loss convergence rates of the hybrid approaches will not be affected even when the i-th observation is included in estimating $\hat{m}_1(X_i)$ and $\hat{m}_0(X_i)$.

$\left(x_1^2 + \cdots + x_{d_x}^2\right)^{1/2}$. Let $K(\cdot) : \mathbb{R}^{d_x} \to \mathbb{R}$ be a kernel function and $h > 0$ be a bandwidth. At each X_i, $i = 1, \ldots, n$, we define the leave-one-out local polynomial coefficient estimators with degree $l \geq 0$ as

$$\begin{aligned}
\hat{\theta}_1(X_i) &= \arg\min_{\theta} \sum_{j \neq i, D_j = 1} \left[Y_j - \theta^T U\left(\frac{X_j - X_i}{h}\right) \right]^2 K\left(\frac{X_j - X_i}{h}\right), \\
\hat{\theta}_0(X_i) &= \arg\min_{\theta} \sum_{j \neq i, D_j = 0} \left[Y_j - \theta^T U\left(\frac{X_j - X_i}{h}\right) \right]^2 K\left(\frac{X_j - X_i}{h}\right),
\end{aligned}$$

where $U\left(\frac{X_j - X_i}{h}\right)$ is the vector with elements indexed by the multi-index s, i.e., $U\left(\frac{X_j - X_i}{h}\right) \equiv \left(\left(\frac{X_j - X_i}{h}\right)^s\right)_{0 \leq |s| \leq l}$.[3] Note that $U(0)$ gives vector $(1, 0, \ldots, 0)^T$. Let $\lambda_{n,1}(X_i)$ be the smallest eigenvalue of $B_1(X_i) \equiv \left(nh^{d_x}\right)^{-1} \sum_{j \neq i, D_j = 1} U\left(\frac{X_j - X_i}{h}\right) U^T\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_i - X_j}{h}\right)$ and $\lambda_{n,0}(X_i)$ be the smallest eigenvalue of $B_0(X_i) \equiv \left(nh^{d_x}\right)^{-1} \sum_{j \neq i, D_j = 0} U\left(\frac{X_j - X_i}{h}\right) U^T\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_i - X_j}{h}\right)$. Accordingly, we construct leave-one-out local polynomial fits for $m_1(X_i)$ and $m_0(X_i)$ by

$$\begin{aligned}
\hat{m}_1(X_i) &= U^T(0)\hat{\theta}_1(X_i) \cdot 1\left\{\lambda_{n,1}(X_i) \geq t_n\right\}, \\
\hat{m}_0(X_i) &= U^T(0)\hat{\theta}_0(X_i) \cdot 1\left\{\lambda_{n,0}(X_i) \geq t_n\right\},
\end{aligned}$$

where t_n is a positive sequence that slowly converges to zero, such as $t_n \propto (\log n)^{-1}$. These trimming rules regularize the regressor matrices of the local polynomial regressions and simplify the proof of the uniform consistency of the local polynomial estimators.

To characterize $\mathcal{P}_m$ in Condition 2.1, we impose the following restrictions.

Assumption E.1.

(Smooth-m) Smoothness of the Regressions: The regression equations $m_1(\cdot)$ and $m_0(\cdot)$ belong to a Hölder class of functions with degree $\beta_m \geq 1$ and constant $L_m < \infty$.[4]

(PX) Support and Density Restrictions on P_X: Let $\mathcal{X} \subset \mathbb{R}^{d_x}$ be the support of P_X. Let $Leb(\cdot)$ be the Lebesgue measure on $\mathbb{R}^{d_x}$. There exist constants $\underline{c}$ and r_0 such that

$$Leb\left(\mathcal{X} \cap B(x, r)\right) \geq \underline{c} Leb(B(x, r)) \quad \forall 0 < r \leq r_0, \forall x \in \mathcal{X}, \tag{E.1}$$

[3] We specify the same degree of polynomial and bandwidth for these two local polynomial regressions only to suppress notational burden.

[4] Let D^s denote the differential operator $D^s \equiv \frac{\partial^{s_1 + \cdots + s_{d_x}}}{\partial x_1^{s_1} \cdots x_{d_x}^{s_{d_x}}}$. Let $\beta \geq 1$ be an integer. For any $x \in \mathbb{R}^{d_x}$ and any $(\beta - 1)$ times continuously differentiable function $f : \mathbb{R}^{d_x} \to \mathbb{R}$, we denote the Taylor expansion polynomial of degree $(\beta - 1)$ at point x by $f_x(x') \equiv \sum_{|s| \leq \beta - 1} \frac{(x' - x)^s}{s!} D^s f(x)$. Let $L > 0$. The Hölder class of functions in $\mathbb{R}^{d_x}$ with degree β and constant $0 < L < \infty$ is defined as the set of function $f : \mathbb{R}^{d_x} \to \mathbb{R}$ that are $(\beta - 1)$ times continuously differentiable and satisfy, for any x and $x' \in \mathbb{R}^{d_x}$, the inequality $|f_x(x') - f(x)| \leq L \|x - x'\|^{\beta}$.

and P_X has the density function $\frac{dP_X}{dx}(\cdot)$ with respect to the Lebesgue measure of $\mathbb{R}^{d_x}$ that is bounded from above and bounded away from zero, $0 < \underline{p}_X \leq \frac{dP_X}{dx}(x) \leq \bar{p}_X < \infty$ for all $x \in \mathcal{X}$.
(Ker) Bounded Kernel with Compact Support: The kernel function $K(\cdot)$ have support $[-1,1]^{d_x}$, $\int_{\mathbb{R}^{d_x}} K(u)du = 1$, and $\sup_u K(u) \leq K_{\max} < \infty$.

Smoothness of the regression equations, Assumption E.1 (Smooth-m), is a standard assumption in the context of nonparametric regressions. Assumption E.1 (PX) is borrowed from Audibert and Tsybakov (2007), and it provides regularity conditions on the marginal distribution of X. Inequality condition (E.1) constrains the shape of the support of X, and it essentially rules out the case where $\mathcal{X}$ has "sharp" spikes, i.e., $\mathcal{X} \cap B(x,r)$ has an empty interior or $Leb(\mathcal{X} \cap B(x,r))$ converges to zero as $r \to 0$ faster than the rate of r^2 for some x in the boundary of $\mathcal{X}$.

Lemma E.4 below shows that when $\mathcal{P}_m$ consists of the data generating processes satisfying Assumption E.1 (Smooth-m) and (PX), Condition 2.1 (m) holds with $\psi_n = n^{\frac{1}{2+d_x/\beta_m}}$, and equation (2.10) in Theorem 2.6 holds with $\tilde{\psi}_n = n^{\frac{1}{2+d_x/\beta_m}} (\log n)^{-\frac{1}{2+d_x/\beta_m}-2}$. The following corollary therefore follows.

Corollary E.1. *Let $\mathcal{P}_m$ consist of data generating processes that satisfy Assumption E.1 (Smooth-m) and (PX). Let $\hat{m}_1(X_i)$ and $\hat{m}_0(X_i)$ be the leave-one-out local polynomial estimators with degree $l = (\beta_m - 1)$, whose kernels satisfy Assumption E.1 (Ker).*

(i) Suppose Assumption 2.1 holds and a bandwidth satisfies $h \propto n^{-\frac{1}{2\beta_m+d_x}}$. Then, it holds

$$\sup_{P \in \mathcal{P}_m \cap \mathcal{P}(M,\kappa)} E_{P^n}\left[W^*_{\mathcal{G}} - W(\hat{G}_{m-hybrid})\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta_m}}\right).$$

(ii) Suppose Assumptions 2.1 and 2.2 hold with margin coefficient $\alpha \in (0,1]$, and a bandwidth satisfies $h \propto \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta_m+d_x}}$. Then, it holds

$$\begin{aligned} &\sup_{P \in \mathcal{P}_m \cap \mathcal{P}_{FB}(M,\kappa,\alpha,\eta)} E_{P^n}\left[W(G^*_{FB}) - W(\hat{G}_{m-hybrid})\right] \\ \leq\ & O\left(n^{-\frac{1+\alpha}{2+d_x/\beta_m}} (\log n)^{\left(\frac{1}{2+d_x/\beta_m}+2\right)(1+\alpha)} \vee n^{-\frac{1+\alpha}{2+\alpha}} \log n\right). \end{aligned}$$

Next, consider the e-hybrid approach. For each $i = 1, \ldots, n$, define a leave-one-out local polynomial propensity score estimator as

$$\begin{aligned} \hat{e}(X_i) &= U^T(0)\hat{\theta}_e(X_i) \cdot 1\{\lambda_n(X_i) \geq t_n\}, \\ \hat{\theta}_e(X_i) &= \arg\min_{\theta} \sum_{j \neq i} \left[D_j - \theta^T U\left(\frac{X_j - X_i}{h}\right)\right]^2 K\left(\frac{X_j - X_i}{h}\right). \end{aligned}$$

We then construct an estimate of individual treatment effect as

$$\hat{\tau}_i = \left[\frac{Y_i D_i}{\hat{e}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{e}(X_i)}\right] \cdot 1\left\{\varepsilon_n \leq \hat{e}(X_i) \leq 1-\varepsilon_n\right\}, \quad 0 < \epsilon_n \leq O\left(n^{-a}\right), a > 0,$$

To ensure Condition 2.1 (e), we assume smoothness of the propensity score function $e(\cdot)$.

Assumption E.2. This assumption is the same as Assumption E.1 except that E.1 *(Smooth-m)* is replaced by

(Smooth-e) Smoothness of the Propensity Score: The propensity score $e(\cdot)$ belongs to a Hölder class of functions with degree $\beta_e \geq 1$ and constant $L_e < \infty$.

Again, Lemma E.4 below shows that $\mathcal{P}_e$ formed by the data generating processes satisfying Assumption E.2, Condition 2.1 (e) holds with $\phi_n = n^{-\frac{1}{2+d_x/\beta_e}}$ and (2.11) with $\tilde{\phi}_n = n^{\frac{1}{2+d_x/\beta_e}} (\log n)^{-\frac{1}{2+d_x/\beta_e}-2}$.

Corollary E.2. *Let $\mathcal{P}_e$ consist of data generating processes that satisfy Assumption E.2 (Smooth-e) and (PX). Let $\hat{e}(X_i)$ be the leave-one-out local polynomial estimator with degree $l = (\beta_e - 1)$, whose kernel satisfy Assumption E.1 (Ker).*

(i) Suppose Assumption 2.1 holds and a bandwidth satisfies $h \propto n^{-\frac{1}{2\beta_e+d_x}}$. Then, it holds

$$\sup_{P \in \mathcal{P}_e \cap \mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^* - W(\hat{G}_{e-hybrid})\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta_e}}\right).$$

(ii) Suppose Assumptions 2.1 and 2.2 hold with margin coefficient $\alpha \in (0,1]$, and a bandwidth satisfies $h \propto \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta_e+d_x}}$. Then, it holds

$$\begin{aligned} &\sup_{P \in \mathcal{P}_e \cap \mathcal{P}_{FB}(M,\kappa,\alpha,\eta)} E_{P^n}\left[W(G_{FB}^*) - W(\hat{G}_{e-hybrid})\right] \\ \leq\ & O\left(n^{-\frac{1+\alpha}{2+d_x/\beta_e}} (\log n)^{\left(\frac{1}{2+d_x/\beta_e}+2\right)(1+\alpha)} \vee n^{-\frac{1+\alpha}{2+\alpha}} \log n\right). \end{aligned}$$

A comparison of Corollaries E.1 and E.2 shows that the rate upper bound of welfare loss differs between the m-hybrid EWM and the e-hybrid EWM approaches when the degree of Hölder smoothness of the regression equations β_m and that of the propensity score β_e are different. For instance, if the propensity score $e(\cdot)$ is smoother than the regression equations of outcome $m_1(\cdot)$ and $m_0(\cdot)$ in the sense of $\beta_e > \beta_m$ and the degree of local polynomial regressions is chosen accordingly, then the rate upper bound of the e-hybrid EWM rule converges faster than that of the m-hybrid EWM rule.

The rest of this section provides formal proofs for validity of Condition 2.1 (m) and (e) for the local polynomial estimators constructed above, when the class of data generating processes $\mathcal{P}_m$ or $\mathcal{P}_e$ is constrained by Assumptions E.1 or E.2. Lemma E.4 shown in Section C.3 proves the main claim. Appendix C.2 collects the preparatory lemmas to prove Lemma E.4.

E.2 Preparatory Lemmas

Let $\mu : \mathbb{R}^{d_x} \to \mathbb{R}$ be a generic notation for a regression equation onto a vector of covariates $X \in \mathbb{R}^{d_x}$. In case of m-hybrid EWM, $\mu(\cdot)$ corresponds to either of $m_1(\cdot)$ or $m_0(\cdot)$. In case of e-hybrid EWM, $\mu(\cdot)$ corresponds to propensity score $e(\cdot)$. We use n to denote the size of the entire sample indexed by $i = 1, \ldots, n$, and denote by $J_i \subset \{1, \ldots, n\}$ a subsample used to estimate $\mu(X_i)$ nonparametrically. Since we consider throughout the leave-one-out regression fits of $\mu(X_i)$, J_i does not include i-th observation. In case of m-hybrid EWM, J_i is either the leave-one-out treated sample $\{j \in \{1, \ldots, n\} : D_j = 1, j \neq i\}$ or the leave-one-out control sample $\{j \in \{1, \ldots, n\} : D_j = 0, j \neq i\}$ depending on $\mu(\cdot)$ corresponds to $m_1(\cdot)$ or $m_0(\cdot)$. Note that, in the m-hybrid case, J_i is random as it depends on a realization of $(D_1, \ldots, D_n)$. When the e-hybrid EWM is considered, J_i is non-stochastic and it is given by $J_i = \{1, \ldots, n\} \setminus \{i\}$. The size of J_i is denoted by n_{J_i}, which is equal to $n_1 - 1$ or $n_0 - 1$ in the m-hybrid case, and is equal to $n - 1$ in the e-hybrid case. With abuse of notations, we use Y_i, $i = 1, \ldots, n$, to denote dependent variable observations and use ξ_i to denote a regression residual, i.e., $Y_i = \mu(X_i) + \xi_i$, $E(\xi_i | X_i) = 0$, holds for all $i = 1, \ldots, n$. For e-hybrid rule, Y_i should be read as the treatment status indicator $D_i \in \{1, 0\}$.

We assume that $\mu(\cdot)$ belongs to a Hölder class of functions with degree $\beta \geq 1$ and constant $0 < L < \infty$. Our generic notation for the leave-one-out local polynomial regression fir for $\mu(X_i)$ with degree $l = (\beta - 1)$ is

$$\begin{aligned} \hat{\mu}_{-i}(X_i) &= U^T(0)\hat{\theta}(X_i) \cdot 1\{\lambda(X_i) \geq t_n\}, \\ \hat{\theta}_{-i}(X_i) &= \arg\min_{\theta} \sum_{j \in J_i} \left[Y_j - \theta^T U\left(\frac{X_j - X_i}{h}\right) \right]^2 K\left(\frac{X_j - X_i}{h}\right), \end{aligned} \tag{E.2}$$

where $U\left(\frac{X_j - X_i}{h}\right)$ is a regressor vector as defined above, $\lambda(X_i)$ is a smallest eigenvalue of $B_{-i}(X_i) \equiv \left(nh^{d_x}\right)^{-1} \sum_{j \in J_i} U\left(\frac{X_j - X_i}{h}\right) U^T\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_i - X_j}{h}\right)$, and t_n is a sequence of trimming constant converging to zero, whose choice is discussed later. The standard least squares calculus shows

$$\hat{\theta}_{-i}(X_i) = B_{-i}(X_i)^{-1} \left(\frac{1}{nh^{d_x}} \sum_{j \in J_i} U\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_j - X_i}{h}\right) \right),$$

so that $\hat{\mu}(X_i)$ can be written as

$$\begin{aligned}\hat{\mu}_{-i}(X_i) &= \left[\sum_{j\in J_i} Y_j \omega_j(X_i)\right] \cdot 1\{\lambda(X_i) \geq t_n\}, \qquad \text{(E.3)}\\ \text{where } \omega_j(X_i) &= \frac{1}{nh^{d_x}} U^T(0) \left[B_{-i}(X_i)\right]^{-1} U\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_j - X_i}{h}\right).\end{aligned}$$

Lemma E.1. *Suppose Assumptions E.1 (PX) and (Ker).*

(i) Conditional on $(X_1, \ldots, X_n)$ such that $\lambda(X_i) > 0$,

$$\begin{aligned}\max_{j\neq i} |\omega_j(X_i)| &\leq c_5 \frac{1}{nh^{d_x}\lambda(X_i)},\\ \sum_{j\in J_i} |\omega_j(X_i)| &\leq \frac{c_5}{nh^{d_x}\lambda(X_i)} \sum_{j\in J_i} 1\left\{(X_j - X_i) \in [-h,h]^{d_x}\right\},\end{aligned}$$

where c_5 is a constant that depends only on β, d_x and $K_{\max}$.

(ii) For any multi-index s such that $|s| \leq (\beta - 1)$, $\sum_{j\in J_i} \left(\frac{X_j - X_i}{h}\right)^s \omega_j(X_i) = 0$.

(iii) Let $\tilde{\lambda}(x)$ be a smallest eigenvalue of $B(x) \equiv \left(nh^{d_x}\right)^{-1} \sum_{j=1}^n U\left(\frac{X_j - x}{h}\right) U^T\left(\frac{X_j - x}{h}\right) K\left(\frac{X_j - x}{h}\right)$ there exist positive constants c_6 and c_7 that depend only on $\underline{c}$, r_0, $\underline{p}_X$, and $K(\cdot)$ such that

$$P^n\left(\left\{\tilde{\lambda}(x) \leq c_6\right\}\right) \leq 2\left[\dim U\right]^2 \exp\left(-c_7 nh^{d_x}\right)$$

holds for all x, P_X-almost surely, at every $n \geq 1$.

Proof. (i) Since $\|U(0)\| = 1$, it holds

$$\begin{aligned}|\omega_j(X_i)| &\leq \frac{1}{nh^{d_x}} \left\|\left[B_{-i}(X_i)\right]^{-1} U\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_j - X_i}{h}\right)\right\|\\ &\leq \frac{K_{\max}}{nh^{d_x}\lambda(X_i)} \left\|U\left(\frac{X_j - X_i}{h}\right) 1\left\{(X_j - X_i) \in [-h,h]^{d_x}\right\}\right\|\\ &\leq \frac{K_{\max} \dim(U)^{1/2}}{nh^{d_x}\lambda(X_i)}\\ &\equiv \frac{c_5}{nh^{d_x}\lambda(X_i)},\end{aligned}$$

for every $1 \leq j \leq n$. Similarly,

$$\begin{aligned}\sum_{j\in J_i} |\omega_j(X_i)| &\leq \frac{K_{\max}}{nh^{d_x}\lambda(X_i)} \sum_{j\in J_i} \left\|U\left(\frac{X_j - X_i}{h}\right)\right\| 1\left\{(X_j - X_i) \in [-h,h]^{d_x}\right\}\\ &= \frac{c_5}{nh^{d_x}\lambda(X_i)} \sum_{j\in J_i} 1\left\{(X_j - X_i) \in [-h,h]^{d_x}\right\}.\end{aligned}$$

(ii) This claim follows from the first order condition for θ in the least square minimization problem in (E.2).

(iii) This lemma is from Equation (6.3, pp. 626) in the proof of Theorem 3.2 in Audibert and Tsybakov (2007), where suitable choices of constant c_6 and c_7 are given in Equation (6.2, pp.625) in Audibert and Tsybakov (2007). □

The next lemma provides an exponential tail bound for the local polynomial estimators. The first statement is borrowed from Theorem 3.2 in Audibert and Tsybakov (2007), and the second statement is its immediate extension.

Lemma E.2. *(i) Suppose Assumption E.1 (PX) and (Ker) hold, and $\mu(\cdot)$ belongs to a Hölder class of functions with degree $\beta \geq 1$ and constant $0 < L < \infty$. Assume J_i is non-stochastic with $n_{J_i} = n - 1$ (e-hybrid case). Then, there exist positive constants c_8, c_9, and c_{10} that depend only on β, d_x, L, $\underline{c}$, r_0, $\underline{p}_X$, and $\bar{p}_X$, such that, for any $0 < h < r_0/\underline{c}$, any $c_8 h^\beta < \delta$, and any $n \geq 2$,*

$$P^{n-1}\left(\left|\hat{\mu}_{-n}(x) - \mu(x)\right| > \delta\right) \leq c_9 \exp\left(-c_{10} n h^{d_x} \delta^2\right),$$

holds for almost all x with respect to P_X, where $P^{n-1}(\cdot)$ is the distribution of $\left\{(Y_i, X_i)_{i=1}^{n-1}\right\}$.

(ii) Suppose Assumptions 2.1 (SO), E.1 (PX), and (Ker) hold, and $\mu(\cdot)$ belongs to a Hölder class of functions with degree $\beta \geq 1$ and constant $0 < L < \infty$. Assume J_i is stochastic (m-hybrid case) with $J_i = \{j \neq i : D_j = d\}$, $d \in \{1, 0\}$. There exist positive constants c_{11}, c_{12}, and c_{13} that depend only on κ, β, d_x, L, $\underline{c}$, r_0, $\underline{p}_X$, and $\bar{p}_X$, such that for any $0 < h < r_0/\underline{c}$, any $c_{11} h^\beta < \delta$, and any $n_{J_n} \geq 1$,

$$P^{n-1}\left(\left|\hat{\mu}_{-n}(x) - \mu(x)\right| > \delta | n_{J_n}\right) \leq c_{12} \exp\left(-c_{13} n_{J_n} h^{d_x} \delta^2\right)$$

holds for almost all x with respect to P_X, where $P^{n-1}(\cdot|n_{J_n})$ is the conditional distribution of $\left\{(Y_i, X_i)_{i=1}^{n-1}\right\}$ given $\sum_{j=1}^{n-1} 1\{D_j = d\}$.

Proof. (i) See Theorem 3.2 in Audibert and Tsybakov (2007).

(ii) Under Assumption 2.1 (SO), the conditional distribution of covariates X given $D = d$, $d \in \{1, 0\}$, has the support $\mathcal{X}$ same as the unconditional distribution P_X, and has bounded density on $\mathcal{X}$, since

$$\frac{\kappa}{1-\kappa}\frac{dP_X}{dx} < \frac{dP_{X|D=d}}{dx} < \frac{1-\kappa}{\kappa}\frac{dP_X}{dx}$$

holds for all $x \in \mathcal{X}$. Therefore, when P_X satisfies Assumption E.1 (PX), the conditional distributions $P_{X|D=d}$, $d \in \{1, 0\}$ also satisfy the support and density conditions analogous to Assumption

E.1 (PX). This implies that, even when we condition on $n_{J_n} = \sum_{j=1}^{n-1} 1\{D_j = d\} \geq 1$, the exponential inequality of (i) in the current lemma is applicable with different constant terms. □

The next lemma concerns an upper bound of the variance of the supremum of centered empirical processes indexed by a class of sets.

Lemma E.3. *Let $\mathcal{B}$ be a countable class of sets in $\mathcal{X}$, and let $\{P_{X,n}(B) : B \in \mathcal{B}\}$ be the empirical distribution based on iid observations, $(X_1, \ldots, X_n)$, $X_i \sim P_X$.*

$$Var\left(\sup_{B\in\mathcal{B}}\{P_{X,n}(B) - P_X(B)\}\right) \leq \frac{2}{n}E\left[\sup_{B\in\mathcal{B}}\{P_{X,n}(B) - P_X(B)\}\right] + \frac{1}{4n}.$$

Proof. In Theorem 11.10 of Boucheron et al. (2013), setting $X_{i,s}$ at the centered indicator function $1\{X_i \in B\} - P_X(B)$, and dividing the inequality of Theorem 11.10 of Boucheron et al. (2013) by n^2 lead to

$$\begin{aligned} Var\left(\sup_{B\in\mathcal{B}}\{P_{X,n}(B) - P_X(B)\}\right) &\leq \frac{2}{n}E\left[\sup_{B\in\mathcal{B}}\{P_{X,n}(B) - P_X(B)\}\right] \\ &\quad + \frac{1}{n}\sup_{B\in\mathcal{B}}\{P_X(B)[1 - P_X(B)]\} \\ &\leq \frac{2}{n}E\left[\sup_{B\in\mathcal{B}}\{P_{X,n}(B) - P_X(B)\}\right] + \frac{1}{4n}. \end{aligned}$$

□

E.3 Main Lemmas and Proofs of Corollaries E.1 and E.2

The next lemma yields Corollaries E.1 and E.2.

Lemma E.4. *Let $\mathcal{P}_\mu$ be a class of joint distributions of (Y, X) such that $\mu(\cdot)$ belongs to a Hölder class of functions with degree $\beta \geq 1$ and constant $0 < L < \infty$, and Assumption E.1 (PX) holds. Let $\hat{\mu}_{-i}(\cdot)$ be the leave-one-out local polynomial fit for $\mu(X_i)$ defined in (E.2), whose kernel function satisfies Assumption E.1 (Ker).*

(i) Then,

$$\sup_{P\in\mathcal{P}_\mu} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\mu}_{-i}(X_i) - \mu(X_i)\right|\right] \leq O(h^\beta) + O\left(\frac{1}{\sqrt{nh^{d_x}}}\right) \tag{E.4}$$

holds. Hence, an optimal choice of bandwidth that leads to the fastest convergence rate of the uniform upper bound is $h \propto n^{-\frac{1}{2\beta+d_x}}$ *and the resulting uniform convergence rate is*

$$\sup_{P\in\mathcal{P}_\mu} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta}}\right).$$

(ii) Let $t_n \propto (\log n)^{-1}$*. Then,*

$$\sup_{P\in\mathcal{P}_\mu} E_{P^n}\left[\left(\max_{1\leq i\leq n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|\right)^2\right] \leq O\left(\frac{h^{2\beta}}{t_n^2}\right)+O\left(\frac{\log n}{nh^{d_x}t_n^2}\right) \tag{E.5}$$

holds. Hence, an optimal choice of bandwidth that leads to the fastest convergence rate of the uniform upper bound is $h \propto \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+d_x}}$ *and the resulting uniform convergence rate is*

$$\sup_{P\in\mathcal{P}_\mu} E_{P^n}\left[\left(\max_{1\leq i\leq n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|\right)^2\right] \leq O\left((t_n)^{-2}\left(\frac{\log n}{n}\right)^{\frac{2}{2+d_x/\beta}}\right).$$

Proof. (i) First, consider the non-stochastic J_i case with $n_{J_i} = (n-1)$ (e-hybrid case). Since observations are iid (hence exchangeable) and the probability law of $\hat{\mu}_{-i}(\cdot)$ does not depend on X_i, it holds

$$\begin{aligned}
E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|\right] &= E_{P^n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right| \\
&= E_{P_X}\left[E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(X_n)-\mu\left(X_n\right)\right||X_n\right]\right] \\
&= \int_{\mathcal{X}} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x)-\mu\left(x\right)\right|\right] dP_X(x) \\
&= \int_{\mathcal{X}}\left[\int_0^\infty P^{n-1}\left(\left|\hat{\mu}_{-n}(x)-\mu\left(x\right)\right|>\delta\right) d\delta\right] dP_X(x),
\end{aligned} \tag{E.6}$$

where $E_{P^{n-1}}[\cdot]$ is the expectation with respect to the first $(n-1)$-observations of (Y_i, X_i). By Lemma E.2 (i), there exist positive constants c_8, c_9, and c_{10} that depend only on β, d_x, L, $\underline{c}$, r_0, $\underline{p}_X$, and $\bar{p}_X$ such that, for any $0 < h < r_0/\underline{c}$, any $c_8 h^\beta < \delta$, and any $n \geq 2$,

$$P^{n-1}\left(\left|\hat{\mu}_{-n}(x)-\mu\left(x\right)\right|>\delta\right) \leq c_9 \exp\left(-c_{10}nh^{d_x}\delta^2\right) \tag{E.7}$$

holds for almost all x with respect to P_X. Hence,

$$\begin{aligned}
\int_{\mathcal{X}}\left[\int_0^\infty P^{n-1}\left(\left|\hat{\mu}_{-n}(x)-\mu\left(x\right)\right|>\delta\right) d\delta\right] dP_X(x) &\leq c_8h^\beta + c_9\int_0^\infty \exp\left(-c_{10}nh^{d_x}\delta^2\right) d\delta \\
&= c_8h^\beta + \frac{c_{14}}{\sqrt{nh^{d_x}}} \\
&= O(h^\beta) + O\left(\frac{1}{\sqrt{nh^{d_x}}}\right)
\end{aligned} \tag{E.8}$$

where $c_{14} = c_9(2c_{10})^{-1/2} \int_0^\infty \left(\delta'\right)^{-1/2} \exp\left(-c_{10}\delta'\right) d\delta' < \infty$. Since the upper bound (E.8) does not depend upon $P \in \mathcal{P}_\mu$, this upper bound is uniform over $P \in \mathcal{P}_\mu$, so the conclusion holds.

Next, consider the stochastic J_i case with $n_{J_i} = \sum_{j \neq i} 1\{D_j = d\}$, where $d \in \{1, 0\}$. we can interpret n_{J_i} as a binomial random variable with parameters $(n-1)$ and π, where $\pi = P(D_i = 1)$ when $\mu(\cdot)$ corresponds to $m_1(\cdot)$ and $\pi = P(D_i = 0)$ when $\mu(\cdot)$ corresponds to $m_0(\cdot)$. In either case, $\kappa < \pi < 1 - \kappa$ by Assumption 2.1 (SO). Let $n \geq 1 + \frac{2}{\pi}$ and $\Omega_{\pi,n} \equiv \left\{\left|\frac{n_{J_n}}{n-1} - \pi\right| \leq \frac{1}{2}\pi\right\} = \left\{\frac{(n-1)\pi}{2} \leq n_{J_n} \leq \frac{3(n-1)\pi}{2}\right\}$. Consider

$$\begin{aligned} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| \cdot 1\{\Omega_{\pi,n}\}\right] &= \sum_{n_{J_n} \in \Omega_{\pi,n}} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| | n_{J_n}\right] P^{n-1}(n_{J_n}) \\ &\leq \max_{n_{J_n} \in \Omega_{\pi,n}} \left\{E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| | n_{J_n}\right]\right\} P^{n-1}(\Omega_{\pi,n}) \\ &\leq \max_{n_{J_n} \in \Omega_{\pi,n}} \left\{E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| | n_{J_n}\right]\right\}. \end{aligned}$$

Since $n_{J_n} \geq \frac{(n-1)\pi}{2} \geq 1$ on $\Omega_{\pi,n}$, Lemma E.2 (ii) implies

$$\begin{aligned} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| | n_{J_n}\right] &\leq \int_{\mathcal{X}} \left[\int_0^\infty P^{n-1}\left(\left|\hat{\mu}_{-n}(x) - \mu(x)\right| > \delta | n_{J_n}\right) d\delta\right] dP_X(x) \\ &\leq c_{11} h^\beta + \frac{c_{15}}{\sqrt{n_{J_n} h^{d_x}}}, \end{aligned}$$

where c_{11} and c_{15} are positive constants that depend only on κ, β, d_x, L, $\underline{c}$, r_0, $\underline{p}_X$, and $\bar{p}_X$. Since $n_{J_n} \geq \frac{(n-1)\pi}{2} \geq \frac{n\pi}{4}$ on $\Omega_{\pi,n}$ for $n \geq 2$, it holds

$$\max_{n_{J_n} \in \Omega_{\pi,n}} \left\{E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| | n_{J_n}\right]\right\} \leq c_{11} h^\beta + \frac{2c_{15}}{\sqrt{\pi n h^{d_x}}}.$$

Accordingly, combined with the Hoeffding's inequality $P^{n-1}\left(\Omega_{\pi,n}^c\right) \leq 2\exp\left(-\frac{\pi^2}{4}n\right)$, we obtain

$$\begin{aligned} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right|\right] &\leq E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| \cdot 1\{\Omega_{\pi,n}\}\right] + M P^{n-1}\left(\Omega_{\pi,n}^c\right) \\ &\leq c_{11} h^\beta + \frac{2c_{15}}{\sqrt{\pi n h^{d_x}}} + 2M \exp\left(-\frac{\pi^2}{4}n\right). \end{aligned}$$

The third term in the right hand side converges faster than the second term, so we have shown

$$\begin{aligned} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n \left|\hat{\mu}_{-i}(X_i) - \mu(X_i)\right|\right] &= \int_{\mathcal{X}} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right|\right] dP_X(x) \\ &\leq O(h^\beta) + O\left(\frac{1}{\sqrt{n h^{d_x}}}\right) \end{aligned}$$

holds for the stochastic J_i case as well.

(ii) Let $\Omega_{\lambda,n}$ be an event defined by $\{\lambda(X_i) \geq t_n,\ \forall i = 1, \ldots, n\}$. On $\Omega_{\lambda,n}$, (E.3) implies

$$\begin{aligned}
\left|\hat{\mu}_{-i}(X_i) - \mu(X_i)\right|^2 &\leq \left|\sum_{j \in J_i} Y_j \omega_j(X_i) - \mu(X_i)\right|^2 \\
&= \left|\sum_{j \in J_i} (\mu(X_j) - \mu(X_i))\,\omega_j(X_i) + \sum_{j \in J_i} \xi_j \omega_j(X_i)\right|^2 \\
&\leq 2\left|\sum_{j \in J_i} (\mu(X_j) - \mu(X_i))\,\omega_j(X_i)\right|^2 + 2\left|\sum_{j \in J_i} \xi_j \omega_j(X_i)\right|^2, \qquad \text{(E.9)}
\end{aligned}$$

where the second line follows from $Y_j = \mu(X_j) + \xi_j$ and $\sum_{j \neq i} \omega_j(X_i) = 0$ as implied by Lemma E.1 (ii). Since $\mu(\cdot)$ is assumed to belong to the Hölder class, Lemma E.1 (ii) and Assumption E.1 (Ker) imply

$$\begin{aligned}
&\left|\sum\nolimits_{j \in J_i} (\mu(X_j) - \mu(X_i))\,\omega_j(X_i)\right|^2 = \left|\sum\nolimits_{j \in J_i} \|X_j - X_i\|^\beta\, \omega_j(X_i)\right|^2 \\
&= \left|\sum\nolimits_{j \in J_i} \|X_j - X_i\|^\beta\, \omega_j(X_i) \cdot 1\left\{(X_j - X_i) \in [-h, h]^{d_x}\right\}\right|^2 \\
&\leq d_x^\beta h^{2\beta} \left|\sum\nolimits_{j \in J_i} |\omega_j(X_i)|\right|^2 \\
&\leq d_x^\beta h^{2\beta} \left(\frac{c_5}{\lambda(X_i)}\right)^2 \left(\frac{1}{nh^{d_x}} \sum\nolimits_{j \in J_i} 1\left\{(X_j - X_i) \in [-h, h]^{d_x}\right\}\right)^2 \\
&\leq c_{16} \frac{h^{2\beta}}{t_n^2} \left(\frac{1}{nh^{d_x}} \sum\nolimits_{j \in J_i} 1\left\{(X_j - X_i) \in [-h, h]^{d_x}\right\}\right)^2,
\end{aligned}$$

where $c_{16} = c_5^2 d_x^\beta$. Under Assumption E.1 (PX) and being conditional on $\Omega_{\lambda,n}$,

$$\begin{aligned}
&\max_{1 \leq i \leq n} \left|\sum\nolimits_{j \in J_i} (\mu(X_j) - \mu(X_i))\,\omega_j(X_i)\right|^2 \leq c_{16} \frac{h^{2\beta}}{t_n^2} \left[\frac{1}{h^{d_x}} \sup_{B \in \mathcal{B}_h} P_{X,n}(B)\right]^2 \\
&\leq c_{16} \frac{h^{2\beta}}{t_n^2} \left[\frac{1}{h^{d_x}} \left(\sup_{B \in \mathcal{B}_h} (P_{X,n}(B) - P_X(B)) + \sup_{B \in \mathcal{B}_h} P_X(B)\right)\right]^2 \\
&\leq c_{16} \frac{h^{2\beta}}{t_n^2} \left[\frac{1}{h^{d_x}} \sup_{B \in \mathcal{B}_h} (P_{X,n}(B) - P_X(B)) + 2^{d_x} \cdot \bar{p}_X\right]^2 \\
&\leq c_{16} \frac{h^{2\beta}}{t_n^2} \left\{\frac{2}{h^{2d_x}} \left[\sup_{B \in \mathcal{B}_h} (P_{X,n}(B) - P_X(B))\right]^2 + 2^{2d_x+1} \cdot \bar{p}_X^2\right\},
\end{aligned}$$

where $\mathcal{B}_h$ is the class of hypercubes in $\mathbb{R}^{d_x}$, $\mathcal{B}_h \equiv \left\{\prod_{k=1}^{d_x} [x_k - h, x_k + h] : (x_1, \ldots, x_{d_x}) \in \mathcal{X}\right\}$, and

the last inequality follows since $(a+b)^2 \leq 2a^2+2b^2$. Accordingly,

$$\begin{aligned} & E_{P^n}\left[\max_{1\leq i\leq n}\left|\sum\nolimits_{j\neq i}\left(\mu\left(X_j\right)-\mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right] \\ \leq\ & c_{17}\frac{h^{2\beta}}{t_n^2}+2c_{16}\frac{h^{2\beta}}{t_n^2}\frac{1}{h^{2d_x}}E_{P^n}\left\{\left[\sup\nolimits_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right]^2\right\} \\ \leq\ & c_{17}\frac{h^{2\beta}}{t_n^2}+4c_{16}\frac{h^{2\beta}}{t_n^2}\frac{1}{h^{2d_x}}\left\{\begin{array}{c} Var\left(\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right) \\ +\left[E_{P^n}\left(\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right)\right]^2 \end{array}\right\}, \end{aligned}$$

where $c_{17}=2^{2d_x+1}c_{16}\bar{p}_X^2$. In order to bound the variance and the squared mean terms in the curly brackets, we apply Lemma E.3 and Lemma A.5 with $\bar{F}=1$ and $\delta=\bar{p}_X\left(2h\right)^{d_x/2}$. Let $v_{\mathcal{B}_h}<\infty$ be the VC-dimension of $\mathcal{B}_h$ that depends only on d_x. For all n satisfying $nh^{d_x}\geq\frac{C_1 v_{\mathcal{B}_h}}{2^{d_x}\bar{p}_X^2}$, we have

$$\begin{aligned} Var\left(\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right) &\leq \frac{2}{n}E_{P^n}\left(\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right)+\frac{1}{4n} \\ &\leq 2^{\frac{d_x}{2}+1}C_2\bar{p}_X\frac{\sqrt{v_{\mathcal{B}_h}h^{d_x}}}{n^{3/2}}+\frac{1}{4n} \quad \text{and} \\ \left[E_{P^n}\left(\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right)\right]^2 &\leq 2^{d_x}C_2^2\bar{p}_X^2\frac{v_{\mathcal{B}_h}h^{d_x}}{n}. \end{aligned}$$

As a result, there exist positive constants c_{18}, and c_{19} that depend only on β, d_x, and $\bar{p}_X$, such that

$$E_{P^n}\left[\max_{1\leq i\leq n}\left|\sum_{j\neq i}\left(\mu\left(X_j\right)-\mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right]\leq c_{17}\frac{h^{2\beta}}{t_n^2}+c_{18}\frac{h^{2\beta}}{t_n^2\left(nh^{d_x}\right)}+c_{19}\frac{h^{2\beta}}{t_n^2\left(nh^{d_x}\right)^{3/2}}$$

holds for all n satisfying $nh^{d_x}\geq\frac{C_1 v_{\mathcal{B}_h}}{2^{d_x}\bar{p}_X^2}$. Since $nh^{d_x}\to\infty$ by the assumption, focusing on the leading term yields

$$\limsup_{n\to\infty}\sup_{P\in P_\mu}E_{P^n}\left[2\max_{1\leq i\leq n}\left|\sum_{j\in J_i}\left(\mu\left(X_j\right)-\mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right]\leq O\left(\frac{h^{2\beta}}{t_n^2}\right). \tag{E.10}$$

In order to bound the second term in the right hand side of (E.9), note first that

$$\begin{aligned} \left|\sum_{j\in J_i}\xi_j\omega_j\left(X_i\right)\right|^2 &\leq \frac{1}{nh^{d_x}\lambda^2(X_i)}\left\|\frac{1}{\sqrt{nh^{d_x}}}\sum_{j\in J_i}\xi_j U\left(\frac{X_j-X_i}{h}\right)K\left(\frac{X_j-X_i}{h}\right)\right\|^2 \\ &\leq \frac{K_{\max}^2}{nh^{d_x}t_n^2}\max_{1\leq k\leq\dim(U)}\eta_{ik}^2 \end{aligned}$$

holds conditional on $\Omega_{\lambda,n}$, where η_{ik}, $1\leq k\leq\dim\left(U\right)$, is the k-th entry of vector

$$\frac{1}{\sqrt{nh^{d_x}}}\sum_{j\in J_i}\xi_j U\left(\frac{X_j-X_i}{h}\right)1\left\{\left(X_j-X_i\right)\in[-h,h]^{d_x}\right\}.$$

Therefore,

$$E_{P^n}\left[\max_{1\leq i\leq n}\left|\sum_{j\in J_i}\xi_j\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right]\leq\frac{K_{\max}^2}{nh^{d_x}t_n^2}E_{P^n}\left[\max_{1\leq i\leq n,1\leq k\leq\dim(U)}\eta_{ik}^2\right]. \tag{E.11}$$

Conditional on $(X_1,\ldots,X_n)$, η_{ik} has mean zero and every summand in η_{ik} lies in the interval, $\left[-\frac{M}{\sqrt{nh^{d_x}}}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\},\frac{M}{\sqrt{nh^{d_x}}}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}\right]$. The Hoeffding's inequality then implies that, for every $1\leq i\leq n$ and $1\leq k\leq\dim(U)$, it holds

$$\begin{aligned}
& P^n\left(|\eta_{ik}|\geq t|X_1,\ldots,X_n\right)\\
\leq\ & 2\exp\left(-\frac{t^2}{\frac{2M^2}{nh^{d_x}}\sum_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}}\right)\\
\leq\ & 2\exp\left(-\frac{t^2}{\frac{2M^2}{nh^{d_x}}\max_{1\leq i\leq n}\sum_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}}\right),\ \ \forall t>0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& E_{P^n}\left[\exp\left(\frac{\eta_{ik}^2}{\frac{2M^2}{nh^{d_x}}\max_{1\leq i\leq n}\sum_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}}\right)|X_1,\ldots,X_n\right]\\
=\ & 1+\int_1^\infty P^n\left(\exp\left(\frac{\eta_{ik}^2}{\frac{2M^2}{nh^{d_x}}\max\limits_{1\leq i\leq n}\sum\limits_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}}\right)\geq t'|X_1,\ldots,X_n\right)dt'\\
=\ & 1+\int_1^\infty P^n\left(|\eta_{ik}|\geq\sqrt{\frac{2M^2}{nh^{d_x}}\max_{1\leq i\leq n}\sum_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}\log t'}|X_1,\ldots,X_n\right)dt'\\
\leq\ & 1+2\int_1^\infty\exp\left(-2\log t'\right)dt'\\
=\ & 1+2\int_1^\infty\left(t'\right)^{-2}dt'\\
=\ & 3
\end{aligned}$$

for all $1\leq i\leq n$ and $1\leq k\leq\dim(U)$. We can therefore apply Lemma 1.6 of Tsybakov (2009) to

bound $E_{P^n}\left[\max_{i,k}\eta_{ik}^2|X_1,\ldots,X_n\right]$,

$$\begin{aligned}
& E_{P^n}\left[\max_{1\leq i\leq n,1\leq k\leq \dim(U)}\eta_{ik}^2|X_1,\ldots,X_n\right] \\
\leq\ & 2M^2\max_{1\leq i\leq n}\left[\frac{1}{nh^{d_x}}\sum_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}\right]\log\left(3\dim\left(U\right)n\right) \\
\leq\ & 2M^2\left[\frac{1}{h^{d_x}}\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)+2^{d_x}\bar{p}_X\right]\log\left(3\dim\left(U\right)n\right).
\end{aligned}$$

By applying Lemma A.5 with $\bar{F}=1$ and $\delta=\bar{p}_X\left(2h\right)^{d_x/2}$, the unconditional expectation of $\max_{i,k}\eta_{ik}^2$ can be bounded as

$$E_{P^n}\left[\max_{1\leq i\leq n,1\leq k\leq \dim(U)}\eta_{ik}^2\right]\leq 2M^2\left[C_2 2^{d_x/2}\bar{p}_X\sqrt{\frac{v_{\mathcal{B}_h}}{nh^{d_x}}}+2^{d_x}\bar{p}_X\right]\log\left(3\dim\left(U\right)n\right) \tag{E.12}$$

for all n such that $nh^{d_x}\geq\frac{C_1 v_{\mathcal{B}_h}}{2^{d_x}\bar{p}_X^2}$. Plugging (E.12) back into (E.11) and focusing on the leading term give

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_\mu}E_{P^n}\left[\max_{0\leq i\leq n}\left|\sum\nolimits_{j\neq i}\xi_j\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right]\leq O\left(\frac{\log n}{nh^{d_x}t_n^2}\right). \tag{E.13}$$

Combining (E.9), (E.10), and (E.13), we obtain

$$\begin{aligned}
& E_{P^n}\left[\max_{1\leq i\leq n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|^2\right] \\
\leq\ & E_{P^n}\left[\max_{1\leq i\leq n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right]+M^2P^n\left(\Omega_{\lambda,n}^c\right) \\
\leq\ & 2E_{P^n}\left[\max_{1\leq i\leq n}\left|\sum\nolimits_{j\neq i}\left(\mu\left(X_j\right)-\mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right] \\
& +2E_{P^n}\left[\max_{1\leq i\leq n}\left|\sum\nolimits_{j\neq i}\xi_j\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right]+M^2P^n\left(\Omega_{\lambda,n}^c\right), \\
=\ & O\left(\frac{h^{2\beta}}{t_n^2}\right)+O\left(\frac{\log n}{nh^{d_x}t_n^2}\right)+M^2P^n\left(\Omega_{\lambda,n}^c\right),
\end{aligned}$$

so the desired conclusion is proven if $P^n\left(\Omega_{\lambda,n}^c\right)$ is shown to converge faster than the $O\left(\frac{\log n}{nh^{d_x}t_n^2}\right)$ term.

To find the convergence rate of $P^n\left(\Omega_{\lambda,n}^c\right)$, consider first the case of non-stochastic J_i. By

applying Lemma E.1 (iii) with the sample size set at $(n-1)$, we have

$$\begin{aligned} P^n\left(\{\lambda\left(X_i\right) \leq c_6, \text{ for some } 1 \leq i \leq n\}\right) &= nP^n\left(\{\lambda\left(X_n\right) \leq c_6\}\right) \\ &= n\int P^n\left(\lambda(X_n) \leq c_6 | X_n\right) dP_X \\ &= n\int P^{n-1}\left(\lambda(x) \leq c_6\right) dP_X(x) \qquad \text{(E.14)} \\ &\leq 2n\left[\dim U\right]^2 \exp\left(-\frac{c_7}{2} n h^{d_x}\right). \end{aligned}$$

For the case of stochastic J_i, by viewing n_{J_i} as a binomial random variable with parameters $(n-1)$ and π with $\kappa < \pi < 1-\kappa$, and recalling that, when P_X satisfies Assumption E.1 (PX), the conditional distributions $P_{X|D=d}$, $d \in \{1,0\}$ also satisfy the support and density conditions stated in Assumption E.1 (PX), we can apply the exponential inequality shown in Lemma E.1 (iii) to bound $P^{n-1}\left(\lambda(x) \leq c_6 | n_{J_n}\right)$. Hence, with $\Omega_{\pi,n} \equiv \left\{\left|\frac{n_{J_n}}{n-1} - \pi\right| \leq \frac{1}{2}\pi\right\} = \left\{\frac{(n-1)\pi}{2} \leq n_{J_n} \leq \frac{3(n-1)\pi}{2}\right\}$ used above, we have

$$\begin{aligned} P^{n-1}\left(\lambda(x) \leq c_6\right) &\leq P^{n-1}\left(\{\lambda(x) \leq c_6\} \cap \Omega_{\pi,n}\right) + P^{n-1}\left(\Omega_{\pi,n}^c\right) \\ &\leq \max_{n_{J_n} \in \Omega_{\pi,n}} P^{n-1}\left(\lambda(x) \leq c_6 | n_{J_n}\right) + P^{n-1}\left(\Omega_{\pi,n}^c\right). \\ &\leq 2\left[\dim U\right]^2 \exp\left(-\frac{c_7\pi}{4} n h^{d_x}\right) + 2\exp\left(-\frac{\pi^2}{4} n\right), \end{aligned}$$

Plugging this upper bound into (E.14) and focusing on the leading term leads to

$$P^n\left(\{\lambda\left(X_i\right) \leq c_6, \text{ for some } 1 \leq i \leq n\}\right) \leq O\left(n\exp\left(-c_7\frac{\pi}{4} n h^{d_x}\right)\right).$$

Hence, in either of the non-stochastic or the stochastic J_i case, since $t_n \leq c_6$ holds for all large n and the obtained upper bounds are uniform over $P \in \mathcal{P}_\mu$, we conclude

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_\mu} E_{P^n}\left[\max_{1\leq i\leq n} \left|\hat{\mu}_{-i}(X_i) - \mu\left(X_i\right)\right|^2\right] \leq O\left(\frac{h^{2\beta}}{t_n^2}\right) + O\left(\frac{\log n}{n h^{d_x} t_n^2}\right) + O\left(n\exp(-n h^{d_x})\right).$$

Since $t_n = (\log n)^{-1}$ by assumption, $O(n\exp\left(-nh^{d_x}\right))$ converges faster than $O\left(\frac{\log n}{n h^{d_x} t_n^2}\right)$, the leading terms are given by the first two terms, $O\left(\frac{h^{2\beta}}{t_n^2}\right) + O\left(\frac{\log n}{n h^{d_x} t_n^2}\right)$. □

Proof of Corollary E.1. By noting the following inequalities,

$$\begin{aligned} E_{P^n}\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\hat{\tau}^m(X_i)-\tau\left(X_i\right)\right|\right] &\leq E_{P^n}\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\hat{m}_1(X_i)-m_1\left(X_i\right)\right|\right] \\ &\quad +E_{P^n}\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\hat{m}_0(X_i)-m_0\left(X_i\right)\right|\right] \\ E_{P^n}\left[\max_{1\leq i\leq n}\left(\hat{\tau}^m(X_i)-\tau\left(X_i\right)\right)^2\right] &\leq 2E_{P^n}\left[\max_{1\leq i\leq n}\left(\hat{m}_1(X_i)-m_1\left(X_i\right)\right)^2\right] \\ &\quad +2E_{P^n}\left[\max_{1\leq i\leq n}\left(\hat{m}_0(X_i)-m_0\left(X_i\right)\right)^2\right], \end{aligned}$$

we obtain the current corollary by applying Lemma E.4. The resulting uniform convergence rate is given by $\psi_n = n^{\frac{1}{2+d_x/\beta_m}}$. When the assumption (2.10) in Theorem 2.6 is concerned, the corresponding rate is given by $\tilde{\psi}_n = \left[\left(\frac{\log n}{n}\right)^{\frac{1}{2+d_x/\beta_m}}(\log n)^2\right]^{-1}$. □

Proof of Corollary E.2. (i) Assume that n is large enough so that $\varepsilon_n \leq \kappa/2$ holds. Given $\hat{e}\left(X_i\right) \in [\varepsilon_n, 1-\varepsilon_n]$, $\hat{\tau}_i^e - \tau_i$ can be expressed as

$$\hat{\tau}_i^e - \tau_i = \frac{Y_i D_i}{e(X_i)}\left[\frac{e(X_i)-\hat{e}(X_i)}{\hat{e}(X_i)}\right] + \frac{Y_i\left(1-D_i\right)}{1-e(X_i)}\left[\frac{e(X_i)-\hat{e}(X_i)}{1-\hat{e}\left(X_i\right)}\right],$$

so

$$|\hat{\tau}_i^e - \tau_i| \leq \frac{M}{\kappa}\cdot\frac{1}{\hat{e}\left(X_i\right)\left(1-\hat{e}\left(X_i\right)\right)}\cdot|\hat{e}(X_i)-e(X_i)|$$

holds. On the other hand, when $\hat{e}\left(X_i\right) \notin [\varepsilon_n, 1-\varepsilon_n]$, $\hat{\tau}_i^e = 0$ and $|\tau_i| \leq \frac{M}{\kappa}$ imply $|\hat{\tau}_i^e - \tau_i| \leq \frac{M}{\kappa}$. Hence, the following bounds are valid,

$$|\hat{\tau}_i^e - \tau_i| \leq \begin{cases} \frac{M}{\kappa}\cdot\frac{4}{\kappa(2-\kappa)}\cdot|\hat{e}(X_i)-e(X_i)| & \text{if } \hat{e}\left(X_i\right) \in \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right], \\ \frac{M}{\kappa}\cdot\frac{1}{\varepsilon_n(1-\varepsilon_n)} & \text{if } \hat{e}\left(X_i\right) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]. \end{cases} \tag{E.15}$$

Hence,

$$\begin{aligned} E_{P^n}\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\hat{\tau}_i^e-\tau_i\right|\right] &= E_{P^n}\left[\left|\hat{\tau}_n^e-\tau_n\right|\right] \\ &\leq \frac{M}{\kappa}\cdot\frac{4}{\kappa\left(2-\kappa\right)}\cdot E_{P^n}\left[\left|\hat{e}(X_n)-e(X_n)\right|\right] \\ &\quad +\frac{M}{\kappa}\cdot\frac{1}{\varepsilon_n\left(1-\varepsilon_n\right)}\cdot P^n\left(\hat{e}\left(X_n\right) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]\right). \end{aligned}$$

By Lemma E.4 (i), $\sup_{P\in\mathcal{P}_e} E_{P^n}\left[|\hat{e}(X_n)-e(X_n)|\right] \leq O(n^{-\frac{1}{2+d_x/\beta_e}})$, so the conclusion follows if $P^n\left(\hat{e}\left(X_n\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\right)$ is shown to converge faster than $O(n^{-\frac{1}{2+d_x/\beta_e}})$. To see this claim is true, note that

$$\begin{aligned} P^n\left(\hat{e}\left(X_n\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\right) &= \int_{\mathcal{X}} P^{n-1}\left(\hat{e}\left(x\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\right)dP_X\left(x\right) \\ &\leq \int_{\mathcal{X}} P^{n-1}\left(\left|\hat{e}\left(x\right)-e(x)\right|\geq\frac{\kappa}{2}\right)dP_X\left(x\right) \\ &\leq c_9\exp\left(-\frac{c_{10}\kappa^2}{4}nh^{d_x}\right) \end{aligned}$$

holds for all n satisfying $c_8h^\beta<\kappa/2$, where the c_8, c_9, and c_{10} are the constants defined in Lemma B.2 (i). Since ε_n is assumed to converge at a polynomial rate, $\frac{1}{\varepsilon_n(1-\varepsilon_n)}P^n\left(\hat{e}\left(X_n\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\right)$ converges faster than $O(n^{-\frac{1}{2+d_x/\beta_e}})$.

(ii) By (E.15), we have

$$\begin{aligned} E_{P^n}\left[\max_{1\leq i\leq n}\left|\hat{\tau}_i^e-\tau_i\right|^2\right] &\leq \left(\frac{4M}{\kappa^2\left(2-\kappa\right)}\right)^2 E_{P^n}\left[\max_{1\leq i\leq n}\left|\hat{e}(X_i)-e(X_i)\right|^2\right] \\ &\quad+\left(\frac{M}{\kappa\varepsilon_n\left(1-\varepsilon_n\right)}\right)^2 P^n\left(\hat{e}\left(X_i\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\text{ for some }1\leq i\leq n\right). \end{aligned} \tag{E.16}$$

By Lemma E.4 (ii), the first term in (E.16) converges at rate $O\left(n^{-\frac{2}{2+d_x/\beta}}(\log n)^{\frac{2}{2+d_x/\beta}+2}\right)$. To find the convergence rate of the second term in (E.16), consider

$$\begin{aligned} & P^n\left(\hat{e}\left(X_i\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\text{ for some }1\leq i\leq n\right) \\ \leq\ & nP^n\left(\hat{e}\left(X_n\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\right) \\ \leq\ & c_9n\exp\left(-\frac{c_{10}\kappa^2}{4}nh^{d_x}\right), \end{aligned}$$

where the last line follows from Lemma B.2 (i). Since ε_n converges at polynomial rate, we conclude the second term in (E.16) converges faster than the first term. □

References

Audibert, J.-Y. and A. B. Tsybakov (2007): "Fast Learning Rates for Plug-in Classifiers," *The Annals of Statistics*, 35, 608–633.

Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration Inequalities, A Nonasymptotic Theory of Independence*, Oxford University Press.

Tsybakov, A. B. (2009): *Introduction to Nonparametric Estimation*, Springer.